Diagnostics → something wrong in your modeling?

- *regression diagnostics*: check model assumptions to suggest further improvement *after* fitting. *The building of an empirical model is an iterative process*. During the process, it is required to check whether *the current fitted model is consistent with data*.
- Q: what assumptions needed to be checked?

prediction / Interpreting β ← 規律 / 隨機 → overall pattern

LNp.7-2

model: $Y=X\beta+\varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$

① ➢ error structure: errors independent, equal variance, normally distributed

② ➢ mean structure: whether $E(Y)=X\beta$ is a correct structure — mainly under-fitting or not

➢ unusual observations: whether *some* observations do not fit the model → individual pattern

- two types of diagnostic techniques: numerical and graphical

When the linear model is correct

- recall (residuals)

surrogate of ε → **Residual** $Y=\hat{Y}+\hat{\varepsilon}$ ("prediction" of ε)

Q: Why cannot get ε from $\hat{\varepsilon}$?

$\dim(\varepsilon)=n$, $\dim(\hat{\varepsilon})=n-p$

$\sum_i \hat{\varepsilon}_i = 0$ (with intercept), $\sum_i \varepsilon_i \neq 0$

➢ prediction: $\hat{Y} = X(X^TX)^{-1}X^TY = HY$, H: hat matrix

➢ residuals: $\hat{\varepsilon} = Y - \hat{Y} = (I-H)Y = (I-H)X\beta + (I-H)\varepsilon \overset{\text{cf.}}{=} (I-H)\varepsilon$

(Note: errors and residuals are different. Q: what difference?)

➢ $\text{var}(\hat{\varepsilon}) = \text{var}((I-H)\varepsilon) = (I-H)^2\sigma^2 = (I-H)\sigma^2 = \sigma^2 \begin{bmatrix} 1-H_{11} & -H_{12} & \cdots & -H_{1n} \\ -H_{21} & 1-H_{22} & & \\ \vdots & & \ddots & \\ -H_{n1} & & & 1-H_{nn} \end{bmatrix}$

H = $[H_{ij}]$

cf. $\text{Var}(\varepsilon)=\sigma^2 I$, $E^*[(I-H)\varepsilon\varepsilon^T(I-H)^T]$

$\Rightarrow$ even though ε is uncorrelated and equal variance, $\hat{\varepsilon}$ may be not

(Note: H depends on X only) → relative variance and correlation of $\hat{\varepsilon}$ are known before observing Y in DOE

$\text{cov}(\hat{\beta})=(X^TX)^{-1}\sigma^2$, $\text{cov}(\hat{Y}) = H\sigma^2$

➢ residuals are useful for detecting lack of fit and checking model assumptions

(Q: Why residuals can do the works?

examine $\hat{\sigma}^2 \overset{?}{=} \sigma^2$

$\hat{\varepsilon}$ for detecting over-fitting? (LNp.6-9)

① $Y = X\beta + \varepsilon = \hat{Y} + \hat{\varepsilon}$ ($X_1(X_1^TX_1)^{-1}X_1^T$)

② true: $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon = (X_1\beta_1 + H_1X_2\beta_2) + ((I-H_1)X_2\beta_2 + \varepsilon) = \hat{Y}_{X_1} + \hat{\varepsilon}^*_{X_1}$ (cf.)

lack of fit → fitted: $Y = X_1\beta_1 + \varepsilon^*$ (11/17)

$X_1[\beta_1 + (X_1^TX_1)^{-1}X_1^TX_2\beta_2]$ check LNp.5-9

Leverage

- leverage: $h_i \equiv H_{ii}$ (Note 1. $\text{var}(\hat{\varepsilon}_i)=(1-h_i)\sigma^2$. Note 2. h_i is known before observing Y)

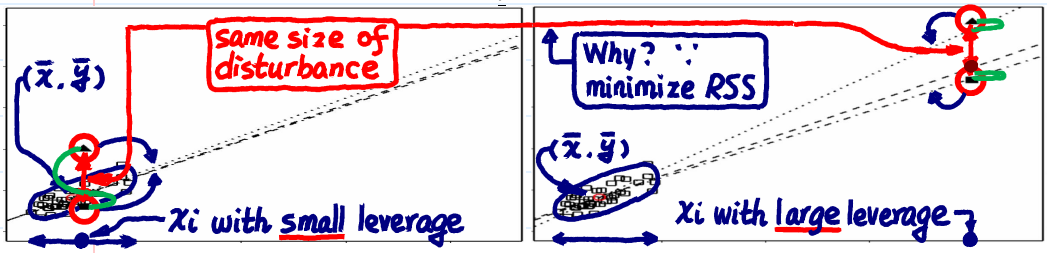
$(x_i, y_i) \leftrightarrow \hat{\varepsilon}_i \leftrightarrow h_i$

x_i whose h_i is large $\Rightarrow$ $\text{var}(\hat{\varepsilon}_i)$ small ($y_i - \hat{y}_i$ small) $\Rightarrow$ fitted model has to force to fit close to y_i ($\hat{y}_i$)

x_i whose h_i is small $\Rightarrow$ $\text{var}(\hat{\varepsilon}_i)$ large $\Rightarrow$ in this x_i, model cannot fit so well ($y_i, \hat{y}_i$ r.v.)

➢ h_i roughly determines how close (x_i, y_i) to the regression surface (i.e., $(x_i, \hat{y}_i)$)

➢ observations with large h_i's should be *paid more attention*. (Q: why?) — $\hat{y}_i$ not so close to y_i



Q: why x_i with large leverage has stronger influence on fit?

➢ for linear model with an intercept, its fitted model must pass the point $(\bar{x}, \bar{y})$

$$E(y) = \beta_0 + \sum_{i=1}^{p-1}\beta_i x_i = \left(\beta_0 + \sum_{i=1}^{p-1}\beta_i \bar{x}_i\right) + \sum_{i=1}^{p-1}\beta_i(x_i - \bar{x}_i) = \beta_0' + \sum_{i=1}^{p-1}\beta_i(x_i - \bar{x}_i)$$

orthogonality also. check LNp.3-7; centering each x; $\bar{y}$; $\mathbf{1} \perp \mathbf{1}$; $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1})$; 力矩 = 力臂 × 作用力

$$\Rightarrow \hat{\beta}_0' = \bar{y} \Rightarrow E(y) = \bar{y} + \sum_{i=1}^{p-1}\hat{\beta}_i(x_i - \bar{x}_i)$$

Q: why is it called leverage?